I4U Submission to NIST SRE 2018: Leveraging from a Decade of Shared Experiences

Kong Aik Lee, Ville Hautamäki, Tomi Kinnunen, Hitoshi Yamamoto, Koji Okabe, Ville Vestman, Jing Huang, Guohong Ding, Hanwu Sun, Anthony Larcher, et al.

I4U Submission to NIST SRE 2018: Leveraging from a Decade of Shared Experiences

Kong Aik Lee[1], Ville Hautamäki[2], Tomi Kinnunen[2], Hitoshi Yamamoto[1], Koji Okabe[1], Ville Vestman[1,2], Jing Huang[3], Guohong Ding[3], Hanwu Sun[4], Anthony Larcher[5], Rohan Kumar Das[6], Haizhou Li[6], Mickael Rouvier[7], Pierre-Michel Bousquet[7], Wei Rao[8], Qing Wang[9], Chunlei Zhang[10], Fahimeh Bahmaninezhad[10], Hector Delgado[11], Jose Patino[11], Qiongqiong Wang[1], Ling Guo[1], Takafumi Koshinaka[1], Jiacen Zhang[1], Koichi Shinoda[1], Trung Ngo Trong[2], Md Sahidullah[2], Fan Lu[3], Yun Tang[3], Ming Tu[3], Kah Kuan Teh[4], Huy Dat Tran[4], Kuruvachan K. George[4], Ivan Kukanov[4], Florent Desnous[5], Jichen Yang[6], Emre Yılmaz[6], Longting Xu[6], Jean-Francois Bonastre[7], Chenglin Xu[8], Zhi Hao Lim[8], Eng Siong Chng[8], Shivesh Ranjan[10], John H.L. Hansen[10], Massimiliano Todisco[11], and Nicholas Evans[11]

[1]NEC Corporation :: Tokyo Institute of Technology, Japan
[2]University of Eastern Finland, Finland :: INRIA, France
[3]JD AI Research and Platform, USA – [4]Institute for Infocomm Research, Singapore
[5]LIUM, France – [6]National University of Singapore, Singapore – [7]LIA, France
[8]Nanyang Technological University, Singapore – [9]Northwest Polytechnic University, China
[10]CRSS, University of Texas at Dallas, USA – [11]EURECOM, France

k-lee@ax.jp.nec.com, villeh@cs.uef.fi, tomi.kinnunen@uef.fi

Abstract

The I4U consortium was established to facilitate a joint entry to NIST speaker recognition evaluations (SRE). The latest edition of such joint submission was in SRE 2018, in which the I4U submission was among the best-performing systems. SRE'18 also marks the 10-year anniversary of I4U consortium into NIST SRE series of evaluation. The primary objective of the current paper is to summarize the results and lessons learned based on the twelve sub-systems and their fusion submitted to SRE'18. It is also our intention to present a shared view on the advancements, progresses, and major paradigm shifts that we have witnessed as an SRE participant in the past decade from SRE'08 to SRE'18. In this regard, we have seen, among others, a paradigm shift from supervector representation to deep speaker embedding, and a switch of research challenge from channel compensation to domain adaptation.

Index Terms: speaker recognition, benchmark evaluation

1. Introduction

The series of speaker recognition evaluations (SRE) conducted by NIST has been a major driving force advancing speaker recognition technology [1, 2]. The basic task is speaker verification: given a segment of speech, decide whether a specified target speaker is speaking in that segment. The SRE in 2018 marks the most recent and ambitious attempt to tackle more realistic tasks [3].

The SRE'18 evaluation set comprises two partitions – *Call-My-Net 2* (CMN2) and *Video-Annotation-for-Speech-Technology* (VAST) – named after the corpora [4, 5] from which the data were derived. For the CMN2 partition, domain mismatch appears to be the major challenge – the *train set* consists of English utterances while the *test set* consists of Tunisian Arabic utterances. For the VAST partition, the major challenge is the *multi-speaker test* scenario, for which an additional diarization module has to be used to determine the target speaker (if any) from a given test segment. This paper presents the technical details of the datasets, sub-system development, and fusion strategy of I4U SRE'18 submission.

In the past decade, I4U participated in five SREs, namely SRE'08, 10, 12, 16, and 18 [6, 7, 8, 9, 10]. Aside from a joint submission, the I4U consortium was formed with a common vision to promote research collaboration and facilitate active exchange of information and experience towards the open evaluation of speaker recognition technology. Along the way we have seen *old* technical challenges were solved, *e.g.*, channel compensation [11, 12, 13], after which researchers have moved on to tackle new challenges, *e.g.*, domain adaptation [14, 15, 16, 17]. SRE18 marks the ten-year anniversary of I4U consortium into NIST SRE series of evaluation. As we set out with the aim to tackle new frontiers in robust speaker recognition, we reckon that it is beneficial looking into past I4U submissions, to share the lessons learned and the insights gained from a decade of I4U experiences.

The paper is organized as follows. Section 2 gives a brief description on SRE'18 dataset, the challenges, and the I4U solutions to deal with them. Then, we present the I4U SRE'18 results in Section 3. Section 4 looks into past I4U submissions. Section 5 concludes the paper.

2. Data and Challenges

Two main challenges of SRE'18 are (i) **domain mismatch** in the CMN2, and (ii) **multi-speaker test** segment in VAST. In this section, we provide a brief description on the CMN2 and VAST data conditions that give rise to the aforementioned challenges and elaborate on the strategy and techniques implemented in I4U sub-systems to deal with them.

Table 1: *List of speech corpora designated as train and development sets for SRE'18 CMN2 and VAST [3].*

Partition	Corpus	Language
CMN2-Train	SRE'04-05-06-08-10-12 Swb-2 Phase I, II, III Swb-Cell Part 1, 2 Fisher 1, 2	English (PSTN)
CMN2-Dev	SRE'18-Dev SRE'18-CMN2-Unlabeled	Tunisian Arabic (PSTN + VOIP)
CMN2-Eval	SRE'18-Eval	
VAST-Train	VoxCeleb1, VoxCeleb2	
VAST-Dev	SRE'18-Dev, SITW-Eval	English (wideband)
VAST-Eval	SRE'18-Eval	

2.1. CMN2 and VAST Partitions

Table 1 shows the list of corpora made available for the fixed-training condition of SRE'18. The train, development and evaluation sets consists of two partitions [3], namely, the *Call-My-Net 2* (CMN2) [4] and *Video-Annotation-for-Speech-Technology* (VAST) [5].

- **CMN2** partition comprises conversational speech in Tunisian Arabic recorded over *voice over internet protocol* (VOIP), in addition to the *public switch telephone network* (PSTN). This is different from Fisher, Switchboard and the Mixer corpora used in previous SREs. Comparing CMN2-Train to the CMN2-Dev and CMN2-Eval sets (see Table 1), two major differences are languages (English versus Tunisian Arabic) and transmission channels (a mix of VOIP and PSTN versus PSTN only). These differences lead to the so-called **domain mismatch** problem, in which the test set does not follow the same distribution as the train.
- **VAST** partition comprises wideband English speech segments extracted from amateur video recordings downloaded from YouTube®. A signature feature of the VAST partition is multi-speaker conversation with considerable background noise. The VoxCeleb [18] and SITW [19] used as the VAST-Train and VAST-Dev, as shown in Table 1, bear the same properties and therefore the same domain.

While it might seem unusual to include two distinct data partitions in a single core task, the setup enables a systematic comparison to past results and system performance on new tasks. In this regard, the CMN2 partition is the continuation of past SREs with new challenges (domain mismatch and lack of labelled in-domain data), while the VAST partition represents a new initiative towards speaker recognition in the wild. See Figure 1. We shall touch upon this point further in Section 4.

2.2. Domain adaptation

A state-of-the-art speaker recognition system consists of a speaker embedding front-end (*e.g.*, i-vector [20], x-vector [21]), followed by a scoring back-end, which is typically implemented with the *probabilistic linear discriminant analysis* (PLDA) [22, 23]. One advantage of the two-stage pipeline is that the same feature extraction and speaker embedding front-end could be used while domain adaptation is accomplished via a transformation on the x-vectors (or i-vectors) [17, 15], or the parameters of the PLDA model [16], to cater for the condition in the anticipated application. The two-stage pipeline design was used for all the twelve sub-systems in I4U SRE'18 submission.

In the case of CMN2, the speaker embedding front-end and PLDA backend are trained on the **out-of-domain** CMN2-Train dataset. Let $\boldsymbol{\phi}$ be the speaker embeddings (*i.e.*, x-vector or i-vector). A PLDA model is given by

$$p\left(\boldsymbol{\phi}\right) = \mathcal{N}\left(\boldsymbol{\phi} | \, \boldsymbol{\mu}, \boldsymbol{\Phi}_{\mathrm{b}} + \boldsymbol{\Phi}_{\mathrm{w}}\right),$$

where $\boldsymbol{\mu}$ is the global mean, $\boldsymbol{\Phi}_{\mathrm{b}}$ and $\boldsymbol{\Phi}_{\mathrm{w}}$ are the between and within-speaker covariance matrices of full rank, respectively. Given SRE18-CMN2-Unlabeled, an unlabelled set of **in-domain** data (see Table 1), the central idea of domain adaptation is to estimate the in-domain between and within-speaker covariance matrices from the in-domain, yet unlabelled, dataset with some helps from out-of-domain covariance matrices. In I4U SRE'18 submission, two unsupervised domain adaptation techniques have been found to be useful, namely, (i) model-level *correlation alignment* with CORAL+ [16], and (ii) Kaldi's PLDA adaptation[1]. We refer the interested reader to [15, 16, 17] and references therein for more details.

2.3. Multi-speaker test segment

The multi-speaker test scenario is not new. It first appeared in NIST SRE'99 [1] where a summed two-channel telephone speech consisting of two speakers was used as the test segment. For the case of SRE'18 VAST partition, there may be several speakers in a test segment. One straightforward solution is to score the entire test segment regardless of other competing speakers. Alternatively, one could use a diarization system to obtain several speaker clusters, score the enrollment segment against all the speaker clusters and select the maximum score. Speaker diarization was explored in Sys. 6 and 7 as shown in Table 2

Following [24], speaker diarization was accomplished using an x-vector PLDA system. Given a VAST test segment, it is first split uniformly into cuts of about 1 second, which are then represented as x-vectors. A matrix of PLDA scores is computed from all the cross-pairs of these x-vectors. The score matrix is used as the affinity matrix in *hierarchical agglomerative clustering* (AHC) where speaker clusters are derived. The number of clusters is determined by an AHC stopping threshold tuned on the SITW set. It is worth mentioning that speaker change point detection which has shown to be critical in reducing the diarization rate seem to be less important in reducing the error rate in speaker verification task.

3. I4U SRE'18 Submission and Results

The sub-system performance is shown in Table 2. Among the twelve sub-systems, eight of them employed x-vector embedding in some form. Notably, Sys. 5 and 6 use attentive pooling layer in the x-vector extractor, while Sys. 10 uses a t-vector embedding trained with a triplet loss [25]. The remaining three sub-systems use i-vector. Comparing the results, x-vector gives a much better performance than i-vector on both CMN2 and VAST. The Kaldi PLDA domain adaptation was the most commonly used strategy. The CORAL+ was also successfully employed resulting in the lowest EER and C_{prim}. Clustering unlabeled set to obtain pseudo-speaker labels was tried in Sys. 3, though no significant difference between clustering and Kaldi adaptation strategy is observed. In terms of the performance on the VAST partition, we observe only slight benefit in using

[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2

Table 2: *Sub-system performance on the NIST SRE'18 evaluation set. Performance is measured in terms of EER and min C_{prim}. We indicate, whether VAST system used diarization (2 systems) and what type of domain adaptation (DA) was utilized (Kaldi PLDA adaptation, CORAL+ or obtaining pseudo labels from the unlabeled set by clustering). Tag 'i' indicates an i-vector system, tag 't' indicates a t-vector, tag 'x' indicates an x-vector, while tag 'x+' indicates an x-vector with attentive pooling.*

			CMN2		VAST	
Sys.	Diar.	DA	EER	C_{prim}	EER	C_{prim}
1 i	N	Kaldi	12.6	0.761	16.8	0.676
2 x	N	Kaldi	11.6	0.759	15.9	0.713
3 x	N	Clust.	8.1	0.549	14.3	0.557
4 x	N	Kaldi	7.5	0.452	**12.1**	**0.543**
5 x+	N	Kaldi	7.9	0.558	15.5	0.637
6 x+	Y	CORAL+	**5.9**	**0.421**	12.7	**0.543**
7 x	Y	Kaldi	7.3	0.491	14.3	0.571
8 x	N	Kaldi	8.1	0.551	14.6	0.601
9 x	N	Kaldi	7.5	0.482	14.3	0.533
10 t	N	Kaldi	10.5	0.678	17.1	0.720
11 i	N	Kaldi	12.4	0.755	18.7	0.700
12 i	N	-	16.4	0.814	21.3	0.788

speaker diarization (Sys. 6 and 7) suggesting a good potential for further improvement.

The scores of the sub-systems were pre-calibrated before fusion. To this end, we apply an affine transformation with simple scaling factor and bias to the scores. The calibrated scores from sub-systems were then combined with a linear fusion. The cross-entropy cost was used for the calibration and fusion with a slight different setting on the effective prior. In this regard, the effective prior was set to 0.5 for score calibration, while an effective prior P_{eff} of 0.005 and 0.05 was used for the fusion for CMN2 and VAST partitions, respectively. Note that the effective priors were set based on those specified in the evaluation plan [3]. The BOSARIS Toolkit [26] was used to perform calibration and fusion. In the primary submission, only subsystems with positive weights were retained. This resulted in 7 subsystems in primary submission of the CMN2 partition (Sys. 3, 4, 6, 7, 9, 10 ,11), and 11 subsystems in the primary submission of the VAST partition (Sys. 1 to 11).

The final submitted fusion system performance is shown in Table 3. In general, the performances on development set and evaluation set agree on the CMN2 partition. On the VAST partition, we notice a large performance gap between development and evaluation sets where the EER increases from 3.70% to 10.18 %. This result reflects the lack of suitable development set for the VAST data. This justifies the use of `SITW` as `VAST-Dev` as shown in Table 1.

4. Past Lessons and Future Outlook

The I4U consortium participated in five SREs in the past decade from SRE'08 to SRE'18. In this section, we look into past I4U results (fusion and single best) to derive insights and to have a glimpse into the current and possible future trends. To start with, we give a brief synopsis and highlight the major challenges in the past SREs.

- SRE'08, 10, and 12 have in common their evaluation sets drawn from the Mixer corpus, or more precisely, different phases of the Mixer corpus [27, 28, 29]. One unique feature of the Mixer corpus is that it consists not only conversational telephone speech (CTS) but also conversational and interview style speech recorded over microphone channel. Among others, one major challenge put forward was cross-channel enrollment and test. This is referred to as the *short2-short3* core task in SRE'08, where the enrollment utterances are either telephone or microphone speech, while the test utterances could be telephone, microphone, or interview speech. SRE'10 followed similar setup except that the core task were split into nine *common conditions* (CCs) corresponding to various combinations of channel (telephone, interview, or microphone) and vocal efforts (low, normal, or high). A larger train set was also provided. SRE'12 has a more complicated setup in which the enrollment utterances were derived from previous SRE'08 and SRE'10, while the test utterances were drawn from previously undisclosed subset of the Mixer corpora. The number of CCs was reduced to five.
- SRE'16 was derived from the *Call-My-Net* corpus [4]. Though the evaluation set is much smaller than that of SRE'12 (few hundreds as opposed to few thousands speakers), SRE'16 posed a new challenge in terms of domain mismatch between the train and evaluation sets. In particular, the train set consists of mainly English speech while the evaluation set was in Tagalog (tgl) and Cantonese (yue). The CMN2 partition of SRE'18 is a continuation of SRE'16 where the same *Call-My-Net* protocol was used to collect speech in Tunisian Arabic [4]. The VAST partition of SRE'18 explores a new direction of data collection from online video [5].

Table 3: *Performance of the primary submissions on the development and evaluation sets.*

CMN2	EER (%)	Min $C_{primary}$	Act $C_{primary}$
Development	4.52	0.277	0.290
Evaluation	5.11	0.362	0.368
VAST	EER (%)	Min $C_{primary}$	Act $C_{primary}$
Development	3.70	0.268	0.300
Evaluation	10.18	0.444	0.550

Table 4 shows the EER of I4U submissions in the past five SREs. Both single-best sub-system and fusion show the same trends. Note that the number of sub-systems used in the fusion varies in each SREs. For SRE'10 and SRE'12, EERs were first computed for each CC and their averages are shown in the table. Figure 1 shows the evolution of EERs on the evaluation set across five past SREs. Strictly speaking, these EERs are not comparable as they were obtained from different evaluation sets. Nevertheless, it is possible to make observations about the general trends.

From SRE'08 to SRE'12, we see that the EER decreases drastically from SRE'08 at 5.90% to 2.23% in SRE'10 and 2.30% in SRE'12. The main theme in these SREs was channel compensation. In this regard, a larger train set benefited significantly channel compensation techniques like *joint factor analysis* (JFA) [12] and *nuisance attribute projection* (NAP) [11] which led to 62% relative EER reduction in SRE'10. In SRE'12, we saw the popularity of i-vector PLDA pipeline [13] as a simpler alternative to JFA where (i) sequence embedding

(i-vector), and (ii) channel compensation and scoring (PLDA) are carried out separately in a pipeline as opposed to a monolithic device. In SRE'12, the EER settled down at similar level as in SRE'10. Compared to its predecessor, the merit of i-vector PLDA is that score normalization is not required. Also shown in Figure 1 are the GMM-SVM (Gaussian mixture model – support vector machine) [11] and GLDS-SVM (generalized linear discriminant sequence kernel SVM), which were two popular technique that use high-dimensional utterance-level representation with SVM.

From SRE'16 to SRE'18 and beyond. We witnessed a rebound in EER with the introduction of CMN evaluation set in SRE'16, which posed a different set of challenges compared to SRE'08-12. Language mismatch and lack of labeled in-domain data are among these challenges. In SRE'18, the EER reduces significantly by 51% from 11.48% to 5.58% on SRE'18 CMN2 partition. Undoubtedly, one major contributor is the x-vector deep speaker embedding method [21, 30]. There is also considerable contribution from unsupervised PLDA adaptation technique as noted in Section 2.2. Another new facet introduced in SRE'18 is the VAST partition. The unconstrained nature of VAST data had proven to be relatively difficult compared to its CMN2 counterpart. We foresee the EER on CMN2 would settle down at around the same level as in SRE'12 when more data is made available. For VAST partition, the difficulty lies at the multi-speaker test segment as noted in Section 2.3. In view of the performance gap between the two partitions, we reckon that new breakthrough in speaker diarization aiming at improving speaker recognition accuracy rather than diarization error is necessary. The forthcoming SRE'19 offers another avenue towards that direction with the use of video information [2].

Large-scale fusion has always been the central stage of I4U submissions. In particular, the I4U submission to SRE'16 encompassed 32 sub-systems, each of them presenting a high-end recognizer involving careful parameter optimization and data engineering. Deploying such massive fusion may be challenging in real use case, reliable fusion indeed plays a key role: it provides a vehicle to solve a common engineering goal, which could not be realistically solved with a single system alone. The SRE'16 fusion result shows that a fairly simple linear fusion improves the performance considerably compared the single-best from 11.48% to 8.59% (see Table 4). Interestingly, in the case of SRE'18 CMN2 we do not observe similar large performance gap, indicating the need for new innovations in the underlying technique. Two other useful points that we can derive from I4U experience are: (i) Score pre-calibration before fusion always help. Notably, it allows classifier selection base on their weights. Classifiers with negative correlation with others will have negative weights and could usually be discarded; (ii) Fusion of fusion (*i.e.*, fusing multiple fused systems) is problematic and should be avoided. The rationale is that it tends to over-fit the Development set.

Channel versus domain mismatch. The notion of channel is used to describe the extrinsic variability imposed on a speech utterance by the acoustic environment, recording device, and the transmission channel. Channel mismatch denotes the inconsistency between the enrollment and test segments in a given trial. For example, a target speaker might be rejected if the channel effects (e.g., enrollment and test utterances of the same speaker but recorded with different devices) is stronger than the speaker characteristic rendered in the utterances. This

[2]https://www.nist.gov/itl/iad/mig/nist-2019-speaker-recognition-evaluation

Table 4: *Performance of I4U fusion and single-best submissions in terms of equal-error-rate (EER) on the evaluation set of SRE'08, 10, 12, 16, and 18 [6, 7, 8, 9, 10].*

	Fusion		Single best
	#sub-systems	EER (%)	EER (%)
SRE'08	7	5.90	6.10
SRE'10	13	2.23	3.55
SRE'12	17	2.30	3.70
SRE'16 CMN	32	8.59	11.48
SRE'18 CMN2	12	5.11	5.86
SRE'18 VAST	12	10.18	12.06

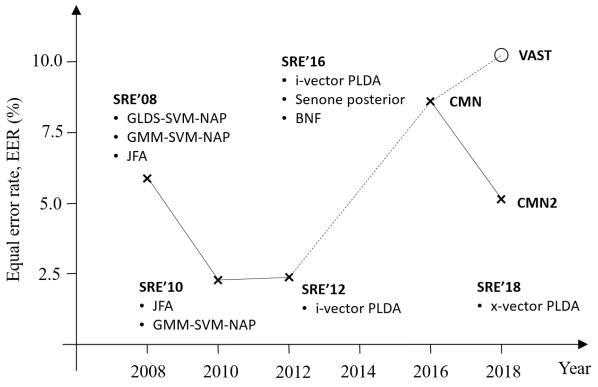

Figure 1: *Progress and performance comparison of I4U submissions from SRE'08 to SRE'18.*

was the main topic in SRE'08, 10, and 12, and had led to the use of channel compensation techniques, like, JFA [12], NAP [11], and PLDA [13]. Domain mismatch, in turn, denotes the inconsistency between Train and Evaluation sets. What this means in the context of SRE'18 CMN2 is that the speaker recognition system was trained with English dataset which is different from from those in which we use the system (i.e., Tunisian Arabic). By domain adaptation, we assume that the channel variability learned from one domain shares some common behaviors in another domain. Simple covariance transformation techniques [15, 16] have shown to work well compared to a much complicated counterpart [31]. This is a topic for future research.

5. Conclusions

This paper presents an overview of the recognition systems and their fusion developed for NIST SRE'18 by I4U consortium. In general, sub-systems that utilized more recent x-vector deep speaker embedding were more successful. On the CMN2 partition, the CORAL+ [16] unsupervised PLDA adaptation technique has shown to be effective. The VAST partition is more difficult compared to the CMN2. One major challenge is the multi-speaker test segment. Marginal improvement was achieved by pre-processing the multi-speaker test segments with a speaker diarization module.

Fusion has always been the center stage of I4U submissions. Comparing the single-best and fusion results in the past SREs from SRE'08 to SRE'18, linear fusion optimized with cross-entropy cost works well. We also found that score pre-calibration helps making classifier selection easier. From SRE'08 to SRE'10 and SRE'12, we observed a significant performance gain in I4U submission due to effective channel com-

pensation techniques (joint factor analysis [12] and PLDA [13]) coupled with a large train set. From SRE'16 to SRE'18, we observed another significant performance gain benefited from the use of deep speaker embedding [21].

6. References

[1] A. Martin and M. Przybocki, "The NIST 1999 speaker recognition evaluationan overview," *Digital Signal Processing*, vol. 10, no. 1, pp. 1 – 18, 2000.

[2] A. F. Martin and J. S. Garofolo, "NIST speech processing evaluations: LVCSR, speaker recognition, language recognition," in *Proc. IEEE Workshop on Signal Processing Applications for Public Security and Forensics*, 2007, pp. 1 – 7.

[3] National Institute of Standards and Technology, "NIST 2018 Speaker Recognition Evaluation Plan," *NIST SRE*, 2018.

[4] K. Jones, S. Strassel, K. Walker, D. Graff, and J. Wright, "Call my net corpus: A multilingual corpus for evaluation of speaker recognition technology," in *Proc. Interspeech 2017*, 2017, pp. 2621–2624. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1521

[5] J. Tracey and S. Strassel, "Vast: A corpus of video annotation for speech technologies," in *Proc. Eleventh International Conference on Language Resources and Evaluation (LREC)*, 2018, pp. 4318–4321.

[6] H. Li, B. Ma, K. A. Lee *et al.*, "The I4U system in NIST 2008 speaker recognition evaluation," in *Proc. ICASSP 2009*, 2009, pp. 4201–4204.

[7] H. Li, B. Ma, H. Sun, K. A. Lee *et al.*, "I4U submission for the 2010 NIST speaker recognition evaluation submission," *NIST SRE 2010 Workshop*, 2010.

[8] R. Saeidi, K. A. Lee, T. Kinnunen *et al.*, "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Proc. Interspeech 2013*, 2013, pp. 1986–1990.

[9] K. A. Lee, V. Hautamaki, T. Kinnunen *et al.*, "The I4U mega fusion and collaboration for NIST speaker recognition evaluation 2016," in *Proc. Interspeech 2017*, 2017, pp. 1328–1332.

[10] ——, "The I4U joint submission for NIST 2018," *NIST SRE 2018 Workshop*, 2018.

[11] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability," in *Proc. IEEE ICASSP*, 2006, pp. 97 – 100.

[12] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448 – 1460, 2007.

[13] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey: Speaker and Language Recognition Workshop*, 2010.

[14] D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey, "Improving speaker recognition performance in the domain adaptation challenge using deep neural networks," in *Proc. IEEE Spoken Language Technology Workshop*, 2014, pp. 378–383.

[15] J. Alam, G. Bhattacharya, and P. Kenny, "Speaker verification in mismatched conditions with frustratingly easy domain adaptation," in *Odyssey: Speaker and Language Recognition Workshop*, 2018.

[16] K. A. Lee, Q. Wang, and T. Koshinaka, "The CORAL+ algorithm for unsupervised domain adaptation of PLDA," in *IEEE ICASSP*, 2019, accepted.

[17] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. AAAI, vol. 6*, 2016, p. 8.

[18] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-950

[19] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database," in *Interspeech 2016*, 2016, pp. 818–822. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1129

[20] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[21] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[22] S. Ioffe, "Probabilistic linear discriminant analysis," in *roceedings of the 9th European Conference on Computer Vision*, 2006.

[23] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV*, 2007, pp. 1–8.

[24] G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *Proceedings of the IEEE Spoken Language Technology Workshop*, 2014.

[25] C. Zhang, K. Koishida, and J. H. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 9, pp. 1633–1644, 2018.

[26] N. Brummer and E. de Villiers, "The bosaris toolkit," Available: https://sites.google.com/site/bosaristoolkit/.

[27] C. Cieri, W. Andrews, J. P. Campbell, G. Doddington, J. Godfrey, S. Huang, M. Liberman, A. Martin, H. Nakasone, M. Przybocki, and K. Walker, "The Mixer and transcript reading corpora: Resources for multilingual, cross-channel speaker recognition research," in *Proc. LREC*, 2006.

[28] L. Brandschain, C. Cieri, D. Graff, A. Neely, and K. Walker, "Speaker recognition: Building the Mixer 4 and 5 corpora," in *Proc. LREC*, 2008.

[29] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition," in *Proc. LREC*, 2010.

[30] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech*, 2018, pp. 2252–2256.

[31] W. Lin, M.-W. Mak, L. Li, and J.-T. Chien, "Reducing domain mismatch by maximum mean discrepancy based autoencoders," in *Proc. Odyssey*, 2018, pp. 162–167.